

Research on the Application of Multi-source Heterogeneous Data Fusion Analysis Technology in Telecom Fraud Detection

Huanxiao Xu^{1*}, Haiwei Yan¹, Chunhua Lin², Hongyun Chen²

¹ School of Yonyou Digital Intelligence, Nantong Institute of Technology, Nantong, China

² School of Information Engineering, Nantong Institute of Technology, Nantong, China

Email Address

1109709073@qq.com (Huanxiao Xu)

*Correspondence: 1109709073@qq.com

Received: 14 January 2026; **Accepted:** 25 February 2026; **Published:** 14 April 2026

Abstract:

The multi-source heterogeneous data fusion analysis method proposed in this paper exhibits significant advantages over general conventional methods to a certain extent. However, to further enhance this advantage, this paper once again adopts the idea of data fusion, incorporating the results from the previously mentioned blacklist/whitelist filtering module, URL feature detection module, and page feature detection module into the judgment criteria of the detection results. Here, a multi-factor comprehensive scoring module that integrates the detection results of all the above modules is proposed. Of course, the more factors that are considered, the better. However, considering practical workload and predictive analysis, this paper only selects the factors proposed in this paper for consideration.

Keywords:

Multi-source Heterogeneous Data Fusion, Phishing Detection, Comprehensive Evaluation of Multiple Factors

1. Introduction

In the era of big data, telecommunications fraud crimes have undergone unpredictable changes, exhibiting significant differences from telecommunications fraud in previous years. Currently, telecommunications fraud relies more on big data technology to collect personal information and then utilizes modern communication technology to conduct non-contact fraud. As a result, both perpetrators and victims are not restricted by geographical boundaries, with cross-regional activities being the main manifestation of the crime. Funds circulate quickly through multiple channels, causing serious losses to people's property. The methods of telecommunications fraud are ever-changing, mainly including telecommunications network phishing, fake lottery fraud, telecommunications password interception fraud, and friend-making fraud. Based on this, this article primarily focuses on one aspect: telecommunication network phishing, which mainly employs means such as impersonating banks, telecommunication companies, websites, etc., through various forms like emails and text messages, to lure victims into logging into websites and disclosing their personal

information and account passwords, subsequently carrying out illegal activities like theft and transfer of funds.

The frequent emergence of phishing websites has not only impacted the development of Internet plus financial services and e-commerce, but also endangered public safety and undermined the confidence of internet users in the internet. To combat the increasingly serious issue of network security, the Anti-Phishing Working Group (APWG) was established internationally in 2003 [1], and the Anti-Phishing Alliance of China was also founded domestically [2]. Furthermore, with the rapid development of artificial intelligence, many researchers have conducted studies related to phishing detection, and research based on machine learning methods has also entered people's field of vision with the advent of big data, being utilized for studying phishing attack detection. In traditional machine learning algorithm-based phishing detection research, Random Forest [3,4] and SVM [5] algorithms are primarily used, as well as Naïve Bayes, C4.5 [6], and logistic regression. With the emergence of new algorithms such as XGBoost [7], they have also been applied to this problem and achieved promising results.

To protect users from phishing attacks, numerous researchers both domestically and internationally have conducted various studies. Most research on phishing website detection falls into four main categories: blacklist and whitelist technology, heuristic detection, visual similarity detection, and machine learning. Among these, blacklist and whitelist detection, heuristic detection, and visual similarity detection all exhibit inherent flaws. The popularity of machine learning algorithms has led most researchers in phishing website detection to focus on machine learning-based approaches. The foundation of detecting phishing websites lies in the presence of sensitive features, which provides the conditions for applying machine learning methods to phishing detection. To achieve accurate and efficient results in machine learning-based phishing website detection models, feature engineering and training classification algorithms are indispensable.

Wang Yuqi et al. [8] proposed a phishing website detection algorithm based on URL linguistic features to address the detection evasion strategies of phishing websites. By analyzing the differences in URLs of phishing websites and legitimate websites across different detection domains, they defined primitives and sensitivities to describe their linguistic features. Initially, similarity detection was performed on the top-level domain names based on primitives. When the similarity fell below a preset threshold, effective subdomain features were selected, and the linguistic features of the subdomains were learned and detected using the random forest algorithm. Experimental results showed that the algorithm achieved an accuracy of 95.6%, with a relatively short system operation time and an average recognition time of less than 1 second. Harinahalli Lokesh Gururaj and Bore Gowda Goutham [9] studied the applicability of machine learning (ML) techniques in identifying phishing attacks and reported their advantages and disadvantages. Specifically, many ML algorithms have been explored as suitable choices for anti-phishing tools. They designed a phishing classification system that extracts features aimed at defeating common phishing detection methods. They also employed numerical representations and conducted comparative studies on classic machine learning techniques, such as random forest, K-nearest neighbors, decision tree, linear SVC classifier, one-class SVM classifier, and wrapper-based feature selection, which incorporates URL metadata and uses this information to determine the legitimacy of a website. Mouad Zouina et al. [10] proposed a novel lightweight phishing detection method based

solely on URLs in 2017. This system uses only six URL features to perform identification. The mentioned features include URL size, number of hyphens, dots, numeric characters, and a discrete variable corresponding to the online status of the IP address in the URL, followed by a similarity index. This system is an SVM (Support Vector Machine) tested on a dataset of 2000 records, including 1000 legitimate URLs and 1000 phishing URLs. Generally speaking, many machine learning methods are based on extracted features. More features may help improve the accuracy of phishing detection. For example, Off-the-Hook [11] extracts more than 200 features. However, too many features can expand the size of the final classifier, ultimately leading to overfitting issues in machine learning methods. To obtain optimal features suitable for the base classifier, many methods have been proposed. Kang Leng Chiew et al. [12] proposed the HEFS (Hybrid Ensemble Feature Selection) method to select the most important features for phishing detection. Almomani Ammar et al. [13] utilized 16 machine learning models, among which 10 semantic features represented the most effective phishing web page detection features extracted from two datasets. Zhu Erzhou et al. [14] introduced the FVV index to assess the significance of each feature in phishing detection. The DT-ANN model proposed in this project falls under the category of machine learning. It employs artificial neural networks to construct the fundamental classifier. To evaluate the importance of different functions on phishing detection performance, the f_Value index is first defined. Subsequently, based on this index, an optimal feature selection algorithm is devised to construct the optimal feature vector for the fundamental neural network classifier.

In addition, in recent years, many scholars have compared existing machine learning algorithms and selected the optimal training classification algorithm. Dželila Mehanović and Jasmin Kevrić [15] solved the problem of phishing website classification. Using the feature selection method of Weka, they employed three classifiers: K-Nearest Neighbors, Decision Tree, and Random Forest. The achieved accuracy was 100%, and the number of features was reduced to seven. Furthermore, as they reduced the number of features, they also decreased the time required to build the model. The time for Random Forest decreased from the initial 2.88s and 3.05s for percentage split and 10-fold cross-validation, respectively, to 0.02s and 0.16s. Bi Qingsong et al. [16] utilized mRMR and RF algorithms to rank features separately; combined the two feature rankings to obtain the final ranking results, and selected the optimal feature subset required for the XGBoost model based on the optimal number of features obtained from experiments; they then used this optimal feature subset to train the XGBoost classification model. Experimental results showed that this method could improve the accuracy of phishing website detection compared to other classification methods, which has practical significance. Ji Tiantian et al. [17] first analyzed and summarized the general process of malicious code attacks. Based on this attack process, they located the empowering attack points and empowering defense points of deep learning, classified the deep learning-assisted attack technologies into five categories, and classified the new deep learning-assisted defense technologies into three categories; secondly, based on the above classification, they reviewed the cutting-edge technologies in malicious code attack and defense research, and conducted in-depth analysis of these technologies from different perspectives such as technical principles, practical feasibility, and development trends. Xu Huanxiao et al. [18] proposed a phishing detection method for laser communication networks based on multi-source heterogeneous data fusion analysis technology. Firstly, they collected phishing features in laser communication networks, fused phishing feature data from different sources and structures using D-S evidence theory, and then designed a

phishing detection model for laser communication networks based on these features using support vector machines. Finally, they conducted simulation experiments, and the results showed that this method could effectively extract and accurately fuse phishing feature data, achieving a phishing detection accuracy of over 95% in laser communication networks, thereby reducing the security risks in laser communication networks.

In summary, through research on existing phishing detection methods both domestically and internationally, it has been found that phishing websites have become increasingly complex and difficult to handle with a single approach in recent years. Given the limitations of various detection methods, more and more researchers are beginning to attempt to improve the detection rate of phishing by improving classification algorithms. This mainly includes: improving a single detection algorithm; parallelizing multiple detection algorithms; and integrating multiple detection algorithms. As a new data fusion technology, the multi-source heterogeneous data fusion method has advantages in the improvement of classification algorithms and has not yet been applied to phishing detection. Therefore, it is completely feasible and highly potential to apply the multi-source heterogeneous data fusion method to the research of phishing detection.

2. Research Content

2.1. Selection of classification algorithm

The so-called classification algorithm refers to the training algorithm of machine learning classification models. Among the numerous training algorithms for machine learning classification models, five representative algorithms are Naive Bayes algorithm, decision tree algorithm, K-Nearest Neighbors (KNN) algorithm, Support Vector Machine (SVM), and neural network algorithm. Common methods for phishing detection involve selecting the optimal solution from a single algorithm. For example, they conducted experimental comparative studies on polynomial Naive Bayes algorithm, SVM, and KNN algorithm based on the page content features and URL features of websites, and concluded that among the three algorithms, the linear SVM classifier had the highest recall and accuracy rates. There are also methods that combine existing machine learning algorithms to select the optimal training classification algorithm compared to individual algorithms. For instance, an algorithm combining K-means and Naive Bayes was compared with using K-means algorithm alone and Naive Bayes algorithm alone, and it was found that the algorithm combining K-means and Naive Bayes was superior.

In summary, after reading previous research and analyses on phishing technique detection, reviewing the current state of research on phishing attack detection, and understanding the advantages and disadvantages of different detection methods, we selected algorithms such as Naive Bayes, KNN, SVM, and Poisson distribution, taking into account the characteristics of machine learning.

2.2. Multi-source heterogeneous data fusion

Given the limitations of various methods, multi-source heterogeneous data fusion has become the future development trend. To find the best improved algorithm, it is necessary to fully understand existing technologies, the features and recognition principles, performance, and adaptability used for phishing website identification, and to design a scheme for multi-information fusion and multi-technology concurrency.

Multi-source heterogeneous data fusion is applied to machine learning algorithms, selecting the most suitable combination method to improve detection effectiveness. The multi-source heterogeneous data fusion methods in this project are diverse, encompassing different data components, content types, and features, including both discrete and mixed data, as well as structured and unstructured data. For example, multi-source heterogeneous machine learning algorithms are used for text classification, image classification, and network link classification, and the classification results are fused to compare and select the optimal solution. After selecting the optimal solution, the Poisson distribution algorithm is incorporated to compare items, thus speculating on whether the more fusion algorithms used, the higher the efficiency. This lays the groundwork for further research in the future.

2.3. Comprehensive evaluation of multiple factors in detection

The multi-source heterogeneous data fusion proposed in this paper exhibits significant advantages over conventional methods to a certain extent. However, to further enhance this advantage, this paper once again adopts the idea of data fusion, incorporating the results of blacklist/whitelist filtering modules, URL feature detection modules, and page feature detection modules involved in phishing detection into the judgment conditions of detection results. Here, a multi-factor comprehensive scoring module that integrates the detection results of all the above modules is proposed. Of course, the more factors considered, the better. However, considering practical workload and predictive analysis, only the factors proposed in this paper are selected for consideration.

3. Technical Approach

The client first sends a user request URL, which is then filtered out by a large portion of webpages through a blacklist. If it is not hit by the blacklist database, it enters the whitelist. After entering the whitelist filtering module, it is compared with the data in the whitelist template database. If it is not hit by the whitelist database, it enters the URL feature recognition module. After entering the URL feature recognition module, URL features are extracted and detected. If they exceed a predetermined threshold, it enters the page feature recognition module. After entering the page feature recognition module, page features are extracted and detected, and the results are compared with conventional methods to reflect the superiority of the new method. Finally, it enters the multi-factor comprehensive scoring module, which synthesizes the experimental results of all previous modules to obtain better detection results.

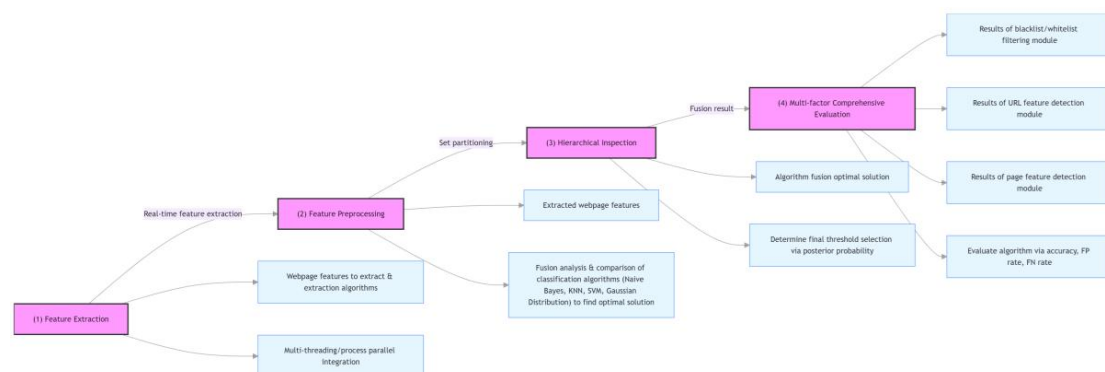


Figure 1. Technology Roadmap.

This project primarily focuses on the research of classification algorithms for the page feature recognition module based on multi-source heterogeneous data fusion. The technical route is illustrated in Figure 1.

4. Conclusions

The multi-source heterogeneous data fusion analysis method proposed in this paper exhibits significant advantages over general conventional methods to a certain extent. However, to further enhance this advantage, this paper once again adopts the idea of data fusion, incorporating the results from the previously mentioned blacklist/whitelist filtering module, URL feature detection module, and page feature detection module into the judgment criteria of the detection results. Here, a multi-factor comprehensive scoring module that integrates the detection results of all the above modules is proposed. Of course, the more factors that are considered, the better. However, considering practical workload and predictive analysis, this paper only selects the factors proposed in this paper for consideration.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by the Nantong Social and People's Livelihood Science and Technology Plan - Guiding Project "Research on the Application of Multi-source Heterogeneous Data Fusion Analysis Technology in Telecommunications Fraud Detection", grant number No.MSZ2024063.

References

- [1] Anti-Phishing Working Group (APWG). APWG 4stQuarter Phishing Activity Trends Report.
- [2] Anti-Phishing Alliance of China (APAC). Brief report on handling phishing websites.
- [3] Yuan, J.D.; Shi, M.H.; Wang, Z.H.; Liu, H.Y.; Li, J.Y. Random pairwise shapelets forest: an effective classifier for time series. *Knowledge and Information Systems*, 2022 (prepublish).
- [4] Zhu, Q.; Lin, G.Y. Research on Phishing Website Detection Method Based on Improved Random Forest Algorithm. *Microelectronics and Computer*, 2019, 4.
- [5] Chittam, S.; Gokaraju, B.; Xu, Z.G.; Sankar, J.; Roy, K. Big Data Mining and Classification of Intelligent Material Science Data Using Machine Learning. *Applied Sciences*, 2021, 11(18).
- [6] Huan, Z.; Jiang, L.X.; Yu, L.J. Class-specific attribute value weighting for Naive Bayes. *Information Sciences*, 2020, 508(C).
- [7] Yang, P.; Zeng, P.; Zhao, G.Z., et al. Phishing website detection method based on logistic regression and XGBoost. *Journal of Southeast University (Natural Science Edition)*, 2019, 2, 1.

- [8] Wang, Y.Q.; Liu, B.W.; Lin Guoyuan. Phishing Website Detection Algorithm Based on URL Language Features. *Computer Engineering and Application*, 2019, 55(24), 84-90.
- [9] Harinahalli, L.G.; BoreGowda, G. Phishing website detection based on effective machine learning approach. *Journal of Cyber Security Technology*, 2021, 5(1).
- [10] Mouad, Z.; Benaceur, O. A novel lightweight URL phishing detection system using SVM and similarity index. *Human-centric Computing and Information Sciences*, 2017.
- [11] Marchal, S.; Armano, G.; Grondahl, T, et al. Off-the-Hook: an efficient and usable client-side phishing prevention application. *IEEE Transactions on Computers*, 2017, 66(10), 1717-1733
- [12] Kang, L.C.; Choon, L.T.; Kok, S.W, et al. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, 2019, 484, 153-166.
- [13] Almomani, A.; Alauthman, M.; Shatnawi, T.M.; Alweshah, M.; Alrosan, A.; Alomoush, W.; Gupta, B.B. Phishing Website Detection With Semantic Features Based on Machine Learning Classifiers: A Comparative Study. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2022, 18(1).
- [14] Zhu, E.Z.; Chen, Y.Y.; Ye, C.C, et al. OFS-NN: an effective phishing websites detection model based on optimal feature selection and neural network. *IEEE ACCESS*, 2019, 99, 1-1
- [15] Dželila, M.; Jasmin, K. Phishing Website Detection Using Machine Learning Classifiers Optimized by Feature Selection. *Traitement du Signal*, 2020, 37(4).
- [16] Bi, Q.; Liang, X.C.; Chen, S.Q. Phishing Website Detection Based on mRMR-RF Feature Selection and XGBoost Model. *Computer Application and Software*, 2020, 37(09), 296-301.
- [17] Ji, T.T.; Fang, B.X.; Cui, X.; Wang, Z.R.; Gan, R.L.; Han, Y.; Yu, W.Q. Research progress on malicious code attack and defense enabled by deep learning. *Chinese Journal of Computers*, 2021, 44(04), 669-695.
- [18] Xu, H.X.; Chen, H.Y.; Wu, F, et al. Laser Communication Network Phishing Detection Based on Multi-source Heterogeneous Data Fusion Analysis Technology. *Journal of Laser*, 2024, 45(03), 219-223. DOI: 10.14016/j.cnki.jgzz.2024.03.219.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)